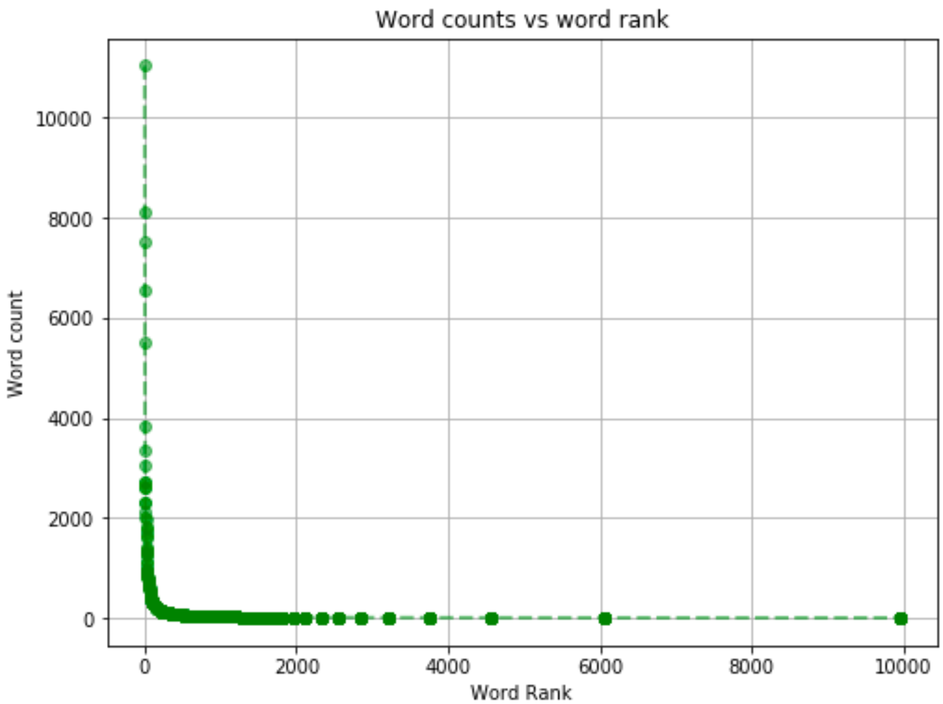
Page 1 Distribution graph

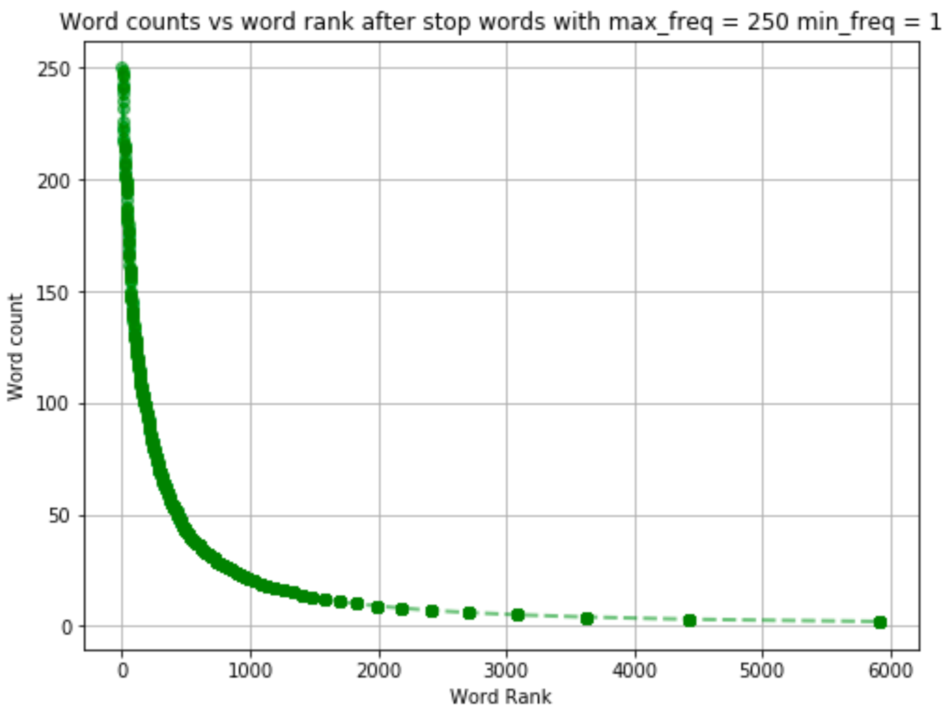


Page 2 Identify the stop words
Try the max frequency threshold = 250 and min. frequency = 1. total 5935 words.
['the', 'and', 'i', 'to', 'a', 'was', 'it', 'of', 'for', 'in', 'my', 'is', 'that', 'they', 'this', 'we', 'you', 'with', 't', 'on', 'not', 'have', 'but', 'had', 'me', 'at', 's', 'so', 'were', 'are', 'be', 'place', 'food', 'there', 'as', 'he', 'if', 'all', 'when', 'out', 'would', 'service', 'get', 'our', 'she', 'back', 'one', 'up', 'time', 'from', 'very', 'an', 'just', 'their', 'here', 'no', 'will', 'great', 'like', 'good', 'go', 'about', 'them', 'or', 'can', 'what', 'your', 'us', 'been', 'do', 'never', 'because', 'only', 'don', 'even', 'after', 'by', 'which', 'did', 'got', 'said', 'more', 'her', 'really', 'told', 'also', 'could', 'some', 'other', 'then', 'went', 've', 'over', 'has', 'well', 'didn', 'again', 'm', 'first', 'best', 'people', 'staff', 'who', 'going', 'came', 'order', 'make', 'any', 'know', 'ordered', 'day', 'ever', 'restaurant', 'how', 'asked', 'off', 'customer', 'am', 'always', 'too', 'come', 'his', 'take', 'took', 'than', 'minutes', 'made', 'experience', 'before', 'try', 'give', 'love', 'car', 'new', 'gai', 'cussed', 'woods', 'woody', 'snowing', 'scold', 'unanswered', 'lord', 'patying', 'sixer', 'miele', 'hacked', 'gerade', 'stabbed', 'polyacrylamide', 'prize', 'clings', 'piling', 'persisted', 'broadstone', 'nigh', 'errors', 'medically', 'ruthless', 'bonuses', 'roadhouse', 'groupie', 'shocks', 'avery', 'erus', 'chino', 'previews','tatum', 'natured', 'poutines', 'learnard', 'oxymoron', 'electricity', 'himmlisch', 'magst', 'scoffed', 'aesthetician', 'geisha', 'accommodation', 'symphony', 'therefore', 'reichhaltig', 'standardized', 'alcoholically', 'inwards', 'intake', 'morally', 'pompadour', 'sweeter', 'concepts', 'zinnias', 'omelets', 'vor', 'couter', 'facade', 'smth', 'cookery', 'hog', 'cutback', 'garments', 'typed', 'coffeeshop', 'cervezas', 'dinosaurs', 'farrrr', 'destined', 'armamnet', 'effective', 'headquarters', 'sickening', 'feedback', 'kochbuch', 'peut', 'partnered', 'bubbles', 'conned', 'lovingly', 'integritythank', 'seasonality', 'fanciest', 'versuchte', 'fig', 'falsch', 'fil', 'songwriter', 'bixi', 'silver', 'ruth', 'horton', 'pleasantries', 'spotless', 'preceded', 'woes', 'spider', 'rt', 'sleepless', 'touts', 'smirk', 'excepting', 'mason', 'cheerfully', 'foundation', 'grapes', 'crowned', 'cadging', 'ministries', 'speedy', 'tempting', 'reserving', 'leaps', 'stripe', 'callous', 'stunnas', 'bacaro', 'millers', 'clarity', 'sherway', 'basketball', 'engagement', 'bitter', ‘wisdom', 'finalement','positively', 'easys', 'sanderson', 'franizkaner', 'hibachi', 'rusty', 'idle', 'sorbets', 'slicers', 'endure', 'tres', 'willingness', 'shampooed', 'chako', 'zzeeks', 'mangoberry', 'alive', 'wholesome', 'crabilicious', 'bagshaw', 'kaya', 'foutent', 'siracha', 'yeishi', 'menudo', 'committing', 'weins', 'object', 'jaser', 'dustbuster', 'welll', 'addict', 'entry', 'singer', 'released', 'camp', 'memberships', 'drunks', 'marvel', 'signatures', 'sweetner', 'loitered', 'prix', 'mcds', 'participate', 'heckled', 'cheaply', 'orleans', 'touches', 'skinnyfats', 'reheating', 'rico', 'bliss', 'foolishly', 'galement', 'chiladas', ‘klein’, honda foremost difara lenders altogether anxiolytic relish avenger saut greene saul sensitivity mamas clem bleh irv santos emptor nlich thevserver skimpy toilette sacrificing stuttgart logic tinge unbuttoned miscommunication advertisements kurma gleaming vais roofer rangers fro twain toning obese regaled cunt commerce contemplated professionally calcium elaborate remembering tzu saltiness characterize doorman portobella toaster oozed craigslist effortlessly bib suppen photographers symmetrical attendais tune maximizing acoustics echoed plaque kilts tilde lassi spectacles emphasis hag vendeurs survival unequivocally rocket otherworldly firmed aussenstehenden tapering bobby froide cambod meisten battlefield crown harassment rotisserie allesamt creep fox witty foe fog indication mildewed picerie binder substituted shifting visitng filipino boiling coronado reathrey statements administer moonen panorama solicitors avail joseph hothead shitiest jang rique underestimated forming notifications verde leider reinventing caromed handily azz refillable brimley moderne minibar regurgitating triangles microwaving actor grammatical truffles hawked outage devon adrianne labbie steelhead irrigation hangovers infants trashcan quartered landeskirche ballet flubbed ndisch bicycles floundering icing prepaid chainiest refinance wharton bustle teas curd skewed settle spiritual rentr boca bagging educate tartare takashi dictates contemplate crumbles brilliant rockbot cette commonly queue accomplished crumbled bearable tasks barbers papas explicitly films ivonne absolument scratches valued believer strain harmonious values fandango mirrors listend cwru nascar matzo monitoring buttered agave elephant grossed loathe mignons provincial lids egotistical insulated sofort sp sw si sm swollen episodes tendency splurge disconnect accordion milked cleanest toro faves…]

```
stop_words_list_max = words['word'][words['count'] > 250].values.tolist() ##threshold = 250
stop_words_list_max += words['word'][words['count'] <=1].values.tolist()
```

Page 3 Distribution graph again

Use the max. freq = 250 and min. freq = 1 to filter stop words.



Page 4 Code snippets

```python
import pandas as pd
import numpy as np
from nltk.corpus import stopwords
df = pd.read_csv('/Users/xinqu/Sandbox/CS498 Applied Machine Learning/HW/HW7/yelp_2k.csv')
df = df[['text', 'stars']]
```

```python
import re
df['clean_text'] = df.text.apply(lambda x: re.sub('[^a-zA-Z]', ' ', x))
from sklearn.feature_extraction.text import CountVectorizer
vectorizer = CountVectorizer(analyzer = "word", tokenizer = None, preprocessor = None,
                             token_pattern = u"(?u)\\b\\w+\\b", stop_words = None, max_features = 100000)
train_data_features = vectorizer.fit_transform(df['clean_text'])
#vectorizer.transform([df.iloc[0]['text']]).toarray()
```

```python
sum_words = train_data_features.sum(axis = 0)
words_freq = [(word, sum_words[0, idx]) for word, idx in vectorizer.vocabulary_.items()]
words_freq = sorted(words_freq, key = lambda x: x[1], reverse =True)
words_freq[:20]
```

```python
words = pd.DataFrame(words_freq, columns = ['word', 'count'])
words['word_rank'] = words['count'].rank(ascending = False)
words.head()
```

	word	count	word_rank
0	the	11041	1.0
1	and	8107	2.0
2	i	7511	3.0
3	to	6565	4.0
4	a	5498	5.0

```python
import matplotlib.pyplot as plt
%matplotlib inline
plt.figure(figsize=(8, 6))
plt.plot(words['word_rank'], words['count'], color='g', linestyle='dashed', marker = 'o',linewidth=2,alpha=0.5)
plt.xlabel('Word Rank')
plt.ylabel('Word count')
plt.title('Word counts vs word rank')
plt.grid()
```

```python
stop_words_list_max = words['word'][words['count'] > 250].values.tolist() ##threshold = 250
stop_words_list_max += words['word'][words['count'] <=1].values.tolist()
```

```python
vec_max = CountVectorizer(analyzer = "word", tokenizer = None, preprocessor = None,
                          token_pattern = u"(?u)\\b\\w+\\b",
                          stop_words = stop_words_list_max, max_features = 100000)
train_data_features_max = vec_max.fit_transform(df['clean_text'])
sum_words_max = train_data_features_max.sum(axis = 0)
words_freq_max = [(word, sum_words_max[0, idx]) for word, idx in vec_max.vocabulary_.items()]
words_freq_max = sorted(words_freq_max, key = lambda x: x[1], reverse =True)
words_max = pd.DataFrame(words_freq_max, columns = ['word', 'count'])
words_max['word_rank'] = words_max['count'].rank(ascending = False)
words_max.head()
```

```python
plt.figure(figsize=(8, 6))
plt.plot(words_max['word_rank'], words_max['count'], color='g', linestyle='dashed',
         marker = 'o',linewidth=2,alpha=0.5)
plt.xlabel('Word Rank')
plt.ylabel('Word count')
plt.title('Word counts vs word rank after stop words with max_freq = 250 min_freq = 1')
plt.grid()
```

```python
#from sklearn.metrics.pairwise import cosine_similarity
max_array = train_data_features_max.toarray() ##bag_of_words threshold = 250
query = vec_max.transform(['Horrible customer service']).toarray()
```

```python
def cos_distance(a, b):
    return 1.0 * np.dot(a, b.T) / (np.linalg.norm(a) * np.linalg.norm(b))
cos_sim = np.zeros(2000,)
for i in range(2000):
    cos_sim[i] = cos_distance(max_array[i], query)
```

```python
df['cosim'] = cos_sim
df_disc = df.sort_values(by = ['cosim'], ascending = False)
for i in range(5):
    text = df_disc.iloc[i]['text']
    dis = df_disc.iloc[i]['cosim']
    print(text + '\n' + 'cosine distance score: ' + str(dis) +'\n')
```

Page 5 Reviews with score

List 5 reviews matching query (top 5 by nearest neighbor with a cos-distance metric)

```
Service was horrible came with a major attitude. Payed 30 for lasagna and was no where worth it. Won't ever be going
back and will NEVER recommend this place. was treated absolutely horrible. Horrible.
cosine distance score: 0.6882472016116852

HORRIBLE HORRIBLE HORRIBLE!!! AVOID AT ALL COSTS!!!

I had some work done at Swing Shift Auto and I was helped by Keith. He was very arrogant and had little time for me.
I just needed new brake discs and pads. I was overcharged, the repairs took TWO DAYS, and when I got home i noticed t
hat the discs had NOT been replaced, only the pads!!!!

TOTAL RIPOFFF!!! NEVER GO HERE, PLEASE!!!
cosine distance score: 0.48038446141526137

Horrible service, horrible customer service, and horrible quality of service!  Do not waste your time or money using
this company for your pool needs.  Dan (602)363-8267 broke my pool filtration system and left it in a nonworking cond
ition.  He will not repair the issue he caused, and told me to go somewhere else.

Save yourself the hassle, there are plenty of other quality pool companies out there.

Take care!
cosine distance score: 0.45226701686664544

I was in there a few weeks ago, the lady who took my order DENE was horrible....not only did she give me the wrong ch
ange but had a terrible attitude and also put in my order wrong not to mention it looked as if she was hungover
If I could give this a negative star I would...what a horrible representation of this place
cosine distance score: 0.42640143271122083

Horrible experience! Got there at 1 am and the front desk worker wasn't there. The lights were turned off so I called
with the after hours phone. After 10 minutes, someone let us in and we stood at the counter and he finally walked up,
then told us we couldn't check in for an hour because the computer was down. Finally got to our room 2 hours later! H
orrible experience!
cosine distance score: 0.35355339059327373
```

Page 6 Query Result

From the above 5 reviews from page 5, the review with cosine score = 0.4527 matches the query well. Among the 5 reviews, only 1 review matches the query well. It contains the query “horrible customer service” and part of the query “horrible service”.

```
Horrible service, horrible customer service, and horrible quality of service!  Do not waste your time or money using
this company for your pool needs.  Dan (602)363-8267 broke my pool filtration system and left it in a nonworking cond
ition.  He will not repair the issue he caused, and told me to go somewhere else.

Save yourself the hassle, there are plenty of other quality pool companies out there.

Take care!
cosine distance score: 0.45226701686664544
```

For the total reviews, by choosing threshold cos-distance = 0.15, there are total 54 reviews matching query. By looking at the cos-distance with text, when cos-distance < 0.15, the text seems to matches only “horrible” in query.

```
1 len(df_disc['cosim'][df_disc['cosim'] >= 0.15])
```

```
54
```

	text	stars	clean_text	cosim
729	Service was horrible came with a major attitud...	1	Service was horrible came with a major attitud...	0.688247
479	HORRIBLE HORRIBLE HORRIBLE!!! AVOID AT ALL COS...	1	HORRIBLE HORRIBLE HORRIBLE AVOID AT ALL COS...	0.480384
90	Horrible service, horrible customer service, a...	1	Horrible service horrible customer service a...	0.452267
1961	I was in there a few weeks ago, the lady who t...	1	I was in there a few weeks ago the lady who t...	0.426401
1763	Horrible experience! Got there at 1 am and the...	1	Horrible experience Got there at am and the...	0.353553
1840	Horrible service....What a mess upon ordering ...	1	Horrible service What a mess upon ordering ...	0.333333
1354	The food is okay but the service is horrible. ...	1	The food is okay but the service is horrible ...	0.333333
1131	I don't understand how people pay money to eat...	1	I don t understand how people pay money to eat...	0.324443
1842	Horrible service! Food was not great either. O...	1	Horrible service Food was not great either O...	0.316228
866	They have no concept of making an authentic sh...	1	They have no concept of making an authentic sh...	0.316228
1808	Rogers ...\n\n1) is over priced\n2) have horri...	1	Rogers is over priced have horrible...	0.316228
887	Horrible service. I was treated so poorly by t...	1	Horrible service I was treated so poorly by t...	0.316228
1646	Would give this place negative stars if I coul...	1	Would give this place negative stars if I coul...	0.316228
1110	This place is horrible. I ask the guy to make ...	1	This place is horrible I ask the guy to make ...	0.316228
1764	Ok. So food is alright. But the service was ho...	1	Ok So food is alright But the service was ho...	0.301511
335	All of the staff here were so kind to me and m...	5	All of the staff here were so kind to me and m...	0.288675
787	Horrible. Worst service ever. Waitstaff was l...	1	Horrible Worst service ever Waitstaff was l...	0.288675
446	Service was ok, food was horrible and there ar...	1	Service was ok food was horrible and there ar...	0.288675
1373	This buffet sucks. Horrible variety, no sides,...	1	This buffet sucks Horrible variety no sides ...	0.277350
1032	Horrible service! I walked in with two bags an...	1	Horrible service I walked in with two bags an...	0.277350
1465	Payless Rent-A-Car is the most horrible place ...	1	Payless Rent A Car is the most horrible place ...	0.274075
1520	HORRIBLE customer service. When you call you'l...	1	HORRIBLE customer service When you call you l...	0.267261
1723	The service is horrible. It's not bad inside, ...	1	The service is horrible It s not bad inside ...	0.267261
838	Horrible snobby service. Took forever. Just wa...	1	Horrible snobby service Took forever Just wa...	0.258199
1335	Horrible attorney. Defend yourself before hir...	1	Horrible attorney Defend yourself before hir...	0.250000
55	Walmart pick up was really horrible.\nI receiv...	1	Walmart pick up was really horrible I receive...	0.242536
161	Used to love this place but recently had the w...	1	Used to love this place but recently had the w...	0.242536
342	everytime we go to this Dennys we have nothing...	1	everytime we go to this Dennys we have nothing...	0.235702
1669	Horrible service! Tried to reservation, I was ...	1	Horrible service Tried to reservation I was ...	0.235702
651	Horrible service and the food is just as bad. ...	1	Horrible service and the food is just as bad ...	0.229416
279	Horrible. Absolutely horrible. Seems like they...	1	Horrible Absolutely horrible Seems like they...	0.227921
642	Horrible horrible place! I don't understand wh...	1	Horrible horrible place I don t understand wh...	0.226455
1403	I was just telling my family about this place ...	1	I was just telling my family about this place ...	0.213201
1114	Thanks for making my shity Monday more shity. ...	1	Thanks for making my shity Monday more shity ...	0.213201
881	The guy who gave me a massage smelled like cig...	1	The guy who gave me a massage smelled like cig...	0.208514
1342	The worst dog grooming I have ever had stay aw...	1	The worst dog grooming I have ever had stay aw...	0.204124
73	Thanksgiving buffet horrible 64$ for some frie...	1	Thanksgiving buffet horrible for some frie...	0.200000
1784	I will never go back to this bar or invite any...	1	I will never go back to this bar or invite any...	0.200000
883	Wish I could give it negative stars\n\nUnbelie...	1	Wish I could give it negative stars Unbelieva...	0.196116
9	Horrible customer service! Been with them ove...	1	Horrible customer service Been with them ove...	0.192450
1852	Horrible please do not go here I went here las...	1	Horrible please do not go here I went here las...	0.188982
732	Wow this food was really horrible. All meats w...	1	Wow this food was really horrible All meats w...	0.188982
1930	I used to love this place, but NEVER order fro...	1	I used to love this place but NEVER order fro...	0.185695
577	Zollie is still not cooperating nor helping to...	1	Zollie is still not cooperating nor helping to...	0.179605
1169	I waited about a week to write this review to ...	1	I waited about a week to write this review to ...	0.178647
1793	Possibly the worst Japanese AYCE I have ever b...	1	Possibly the worst Japanese AYCE I have ever b...	0.174667
72	For a place that has been in business for year...	1	For a place that has been in business for year...	0.174078
801	I have been here multiple times with my friend...	1	I have been here multiple times with my friend...	0.171499
253	Absolutely horrible first experience. Ordered ...	1	Absolutely horrible first experience Ordered ...	0.164399
340	Went in to return an item the greeter was frie...	1	Went in to return an item the greeter was frie...	0.162221
1705	The menu is extremely limited and the service ...	1	The menu is extremely limited and the service ...	0.160128
652	I also wish I could give them zero stars. \n\n...	1	I also wish I could give them zero stars Pl...	0.158114
1201	Eat here at your own risk. This place is horr...	1	Eat here at your own risk This place is horr...	0.156174
245	Worst hotel in Vegas, I'd rather sleep in my c...	1	Worst hotel in Vegas I d rather sleep in my c...	0.154303

Page 7 Accuracy with threshold 0.5

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
X_train, X_test, y_train, y_test = train_test_split(train_data_features_max, df['stars'],
                                                    test_size = 0.1, random_state = 42)
```

```
import warnings
warnings.filterwarnings('ignore')
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression(class_weight = 'balanced')
logreg.fit(X_train, y_train)
threshold = 0.5
predicts = np.where(logreg.predict_proba(X_test)[:, 1] > threshold, 5, 1)
```

```
accuracy_score(np.where(logreg.predict_proba(X_train)[:, 1] > threshold, 5, 1), y_train)
###train set accuracy score
```

```
0.9994444444444445
```

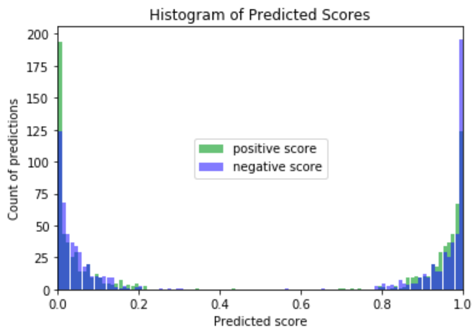
```
accuracy_score(predicts, y_test) ###test set accuracy score
```

```
0.92
```

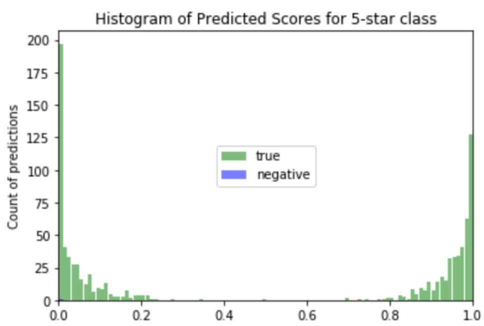
Page 8 Predicted scores

```
pos_train = logreg.predict_proba(X_train)[:, 1]
            [np.where(y_train[logreg.predict_proba(X_train)[:, 1] > threshold]== 5)]
neg_train = logreg.predict_proba(X_train)[:, 0]
            [np.where(y_train[logreg.predict_proba(X_train)[:, 1] <= threshold]== 1)]
```

```
plt.hist(pos_train, bins = 100, color = 'g', alpha = 0.5, density = False, rwidth = 0.9, label = 'positive score')
plt.hist(neg_train, bins = 100, color = 'b', alpha = 0.5, density = False, rwidth = 0.9, label = 'negative score')
plt.xlim(0, 1.0)
plt.xlabel('Predicted score')
plt.ylabel('Count of predictions')
plt.title('Histogram of Predicted Scores')
plt.legend(loc = 'center')
plt.show()
```



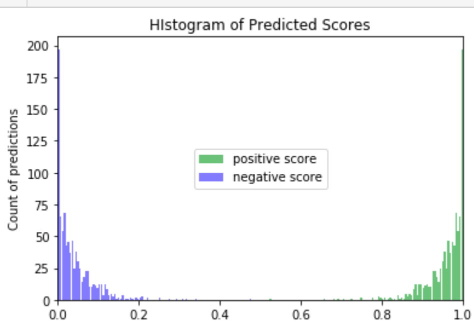

```
pos_train_true = logreg.predict_proba(X_train)[:, 1][np.where(y_train[logreg.predict_proba(X_train)[:, 1] > threshold]== 5)]
pos_train_neg = logreg.predict_proba(X_train)[:, 1][np.where(y_train[logreg.predict_proba(X_train)[:, 1] > threshold]== 1)]
plt.hist(pos_train_true, bins = 100, color = 'g', alpha = 0.5, density = False, rwidth = 0.9, label = 'true')
plt.hist(pos_train_neg, bins = 100, color = 'b', alpha = 0.5, density = False, rwidth = 0.9, label = 'negative')
plt.xlim(0, 1.0)
plt.xlabel('Predicted score')
plt.ylabel('Count of predictions')
plt.title('Histogram of Predicted Scores for 5-star class')
plt.legend(loc = 'center')
plt.show()
```



Since there is only 1 False label in training data, the histogram of false label cannot show in the figure.

```
pos_score = logreg.predict_proba(X_train)[:, 1][np.where(logreg.predict_proba(X_train)[:, 1] > threshold)]
neg_score = logreg.predict_proba(X_train)[:, 0][np.where(logreg.predict_proba(X_train)[:, 0] <= threshold)]
```

```
plt.hist(pos_score, bins = 100, color = 'g', alpha = 0.5, density = False, rwidth = 0.9, label = 'positive score')
plt.hist(neg_score, bins = 100, color = 'b', alpha = 0.5, density = False, rwidth = 0.9, label = 'negative score')
plt.xlim(0, 1.0)
plt.xlabel('Predicted score')
plt.ylabel('Count of predictions')
plt.title('HIstogram of Predicted Scores')
plt.legend(loc = 'center')
plt.show()
```



Page 9 Accuracy again and curve

```
threshold_n = 0.4
predicts_n = np.where(logreg.predict_proba(X_test)[:, 1] > threshold_n, 5, 1)
accuracy_score(np.where(logreg.predict_proba(X_train)[:, 1] > threshold_n, 5, 1), y_train)
###train set accuracy score
```

```
0.9983333333333333
```

```
accuracy_score(predicts_n, y_test) ###test set accuracy score
```

```
0.93
```

Reason for choosing threshold = 0.4: the accuracy score for testing data reaches maximum of 0.93 while the accuracy score for training data still maintains at a high value. Besides, the intersection between positive score and negative score is pretty small, which means false positive rate is small while true positive rate maintains high value.

```
test = np.arange(0.1, 1.0, 0.1)
train_score = []
test_score = []
for i in test:
    predicts = np.where(logreg.predict_proba(X_test)[:, 1] > i, 5, 1)
    train_score.append(accuracy_score(np.where(logreg.predict_proba(X_train)[:, 1] > i, 5, 1), y_train))
    ###train set accuracy score
    test_score.append(accuracy_score(predicts, y_test)) ###test set accuracy score
```

```
train_score, test_score
```

```
([0.9344444444444444,
  0.985,
  0.9955555555555555,
  0.9983333333333333,
  0.9994444444444445,
  1.0,
  0.9977777777777778,
  0.9905555555555555,
  0.94],
 [0.86, 0.885, 0.915, 0.93, 0.92, 0.895, 0.89, 0.86, 0.795])
```

Page 10 Best threshold

```
fpr[np.argmax(tpr)]
```

```
0.22448979591836735
```

From the ROC curve, true positive rate reaches around to 1 when false positive rate is around 0.25. From the above code, the threshold is about 0.22 that minimizes false positive rate while maximizing true positive rate.